HUMAN ACTION RECOGNITION USING ADAPTIVE HIERARCHICAL DEPTH MOTION MAPS AND GABOR FILTER

Hong Liu, Qinqin He, Mengyuan Liu

Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University
hongliu@pku.edu.cn, heqinqin@sz.pku.edu.cn, liumengyuan@pku.edu.cn

ABSTRACT

Depth motion maps (DMMs) have shown effectiveness in human action recognition, however, they lose the temporal information and suffer from intra-class variations caused by action speed variations. To address these challenges, we propose a novel method for human action recognition. Firstly, Adaptive Hierarchical Depth Motion Maps (AH-DMMs) are calculated over temporal hierarchical windows of video sequences to capture the temporal information. Moreover, adaptive windows and steps are employed to ensure that AH-DMMs are robust to motion speed variations. Then, Gabor filter is adopted to encode the texture information of AH-DMMs, generating compact and discriminative action representations. Finally, the representations serve as the input of collaborative representation classifier (CRC). Experimental results on public benchmark MSRAction3D dataset and DHA dataset demonstrate the superiority of the proposed method over the state-of-the-art depth-based action recognition approaches.

Index Terms— Human action recognition, Depth motion maps, Temporal information

1. INTRODUCTION

Human action recognition has been widely applied in a number of real-world applications, e.g., video analysis [1], human-computer interaction [2], and smart surveillance [3]. Though significant progress has been made in past decades [4, 5], recognizing actions is still a quite challenging task due to the inherent limitations of the traditional data source, such as the variations in the lighting conditions, self-occlusions and cluttered backgrounds. With the release of low-cost and easy-operation depth sensor (e.g., Microsoft Kinect), it has become feasible to capture depth information in real-time. Compared with RGB cameras, depth sensors provide the 3D information of human body, which makes it much easier to recover postures and recognize actions.

This work is supported by National High Level Talent Special Support Program, National Natural Science Foundation of China (NSFC, No.61340046, 61673030, U1613209), Specialized Research Fund for the Doctoral Program of Higher Education (No.20130001110011), Natural Science Foundation of Guangdong Province (No.2015A030311034), Scientific Research Project of Guangdong Province (No.2015B010919004).

Relation to prior work: Based on depth images, various representations have been developed for action recognition, such as hyper-surface normals [6, 7], cloud points [8–10], skeleton joints [11, 12] and depth motion maps (DMMs) [13]. Our work is based on DMMs since they can effectively capture the motion and shape clues of human actions. DMMs were proposed by Yang *et al.* [13], they projected depth maps onto three orthogonal planes and accumulated global actions through entire video sequences to generate DMMs, then histogram of oriented gradient (HOG) features were computed from DMMs as the representation of the sequences. Chen *et al.* [14] modified the DMMs by omitting the threshold for real-time action recognition, and they used local binary patterns (LBP) to characterize texture information of DMMs in [15]. More recently, Wang *et al.* [16] calculated three DMMs and each DMM served as an input to a Deep Convolutional Neural Network (ConvNet) for classification. However, in all of above works, DMMs were calculated from the entire video sequence and the temporal information of an action may be lost. Therefore, it is difficult to recognize two actions with similar movements but reverse temporal orders, such as actions "*stand up*" and "*sit down*". And DMMs cannot adapt to motion speed variations, which leads to intra-class variations.

In this paper, we propose an effective method to address above problems. First, a novel descriptor named as Adaptive Hierarchical Depth Motion Maps (AH-DMMs) is proposed to capture the temporal information of actions. The AH-DMMs are calculated over multi-size temporal hierarchical windows, therefore they encode more details of motion and shape information which are lost in DMMs. Meanwhile, by using motion energy based segmentation strategy, adaptive windows and steps are generated, making our AH-DMMs robust to action speed variations. Second, Gabor features encoding the texture information of AH-DMMs are extracted to further improve the discriminative ability of our descriptors. Third, after reducing dimensions by Principle Component Analysis (PCA), the final representations are classified by l_2-regularized CRC. The proposed method can not only depict the motion and shape information, but also take temporal order and action speed variations into consideration. Our method is evaluated on two benchmark datasets and achieves superior performance over the state-of-the-art approaches.

2. THE PROPOSED METHOD

2.1. Adaptive Hierarchical DMMs

DMMs [15] can effectively capture the motion and shape information of human actions. They are generated by projecting the depth frames onto three orthogonal planes and accumulating the difference between projected maps through the entire sequence. Given a depth video sequence with N frames, DMMs can be computed as follows:

$$DMM_{\{f,s,t\}} = \sum_{i=2}^{N} \left| map^{i}_{\{f,s,t\}} - map^{i-1}_{\{f,s,t\}} \right| \quad (1)$$

where $map^{i}_{\{f,s,t\}}$ is the projected map of the i-th depth map on front view, side view and top view. DMMs lose the temporal information of action sequences and cannot adapt to variations of motion speed. To solve these problems, we propose Adaptive Hierarchical Depth Motion Maps (AH-DMMs), which are computed over a series of temporal hierarchical windows to preserve temporal information. In addition, to make our descriptors robust to motion speed, the adaptive windows are selected based on motion energy. The motion energy $ME(i)$ of the i-th frame can be calculated according to [6]. It is modified by removing the threshold for better computational efficiency, which is given as:

$$ME(i) = \sum_{v=1}^{3} \sum_{j=1}^{i-1} \text{num}\left(\left| map_v^{j+1} - map_v^{j} \right| \right) \quad (2)$$

where $\text{num}(\cdot)$ returns the number of non-zero elements in a binary map; $i = 2, ..., N$; $v = \{1, 2, 3\}$ refers to three projection views, respectively. $ME(i)$ reflects the accumulated motion energy form the first frame to i-th frame, $ME(1)$ =0.

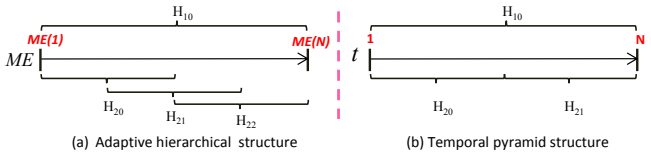


Fig. 1: Comparison of two structures. H_{lm} is the m-th hierarchy in l-th level. (Fig.1 shows only two levels)

The ME of an action sequence is then subdivided into different segments according to the adaptive hierarchical structure, shown in Fig.1(a). In the structure, the level 1 includes only one window which covers $ME(N)$ motion energy. In level 2, each window covers $\frac{1}{2}ME(N)$ motion energy, and the step from one window to the next is half of the window size. Therefore, in level l, each window size W_l and step length S_l can be computed as follows:

$$W_l = (\frac{1}{2})^{l-1} ME(N) \qquad S_l = \frac{1}{2} W_l \quad (3)$$

After dividing ME of the sequence, these segments' corresponding frame indices are used to partition the video sequence. Fig.2 is a specific example of generating AH-DMMs with three levels. The ME is normalized to [0,1],

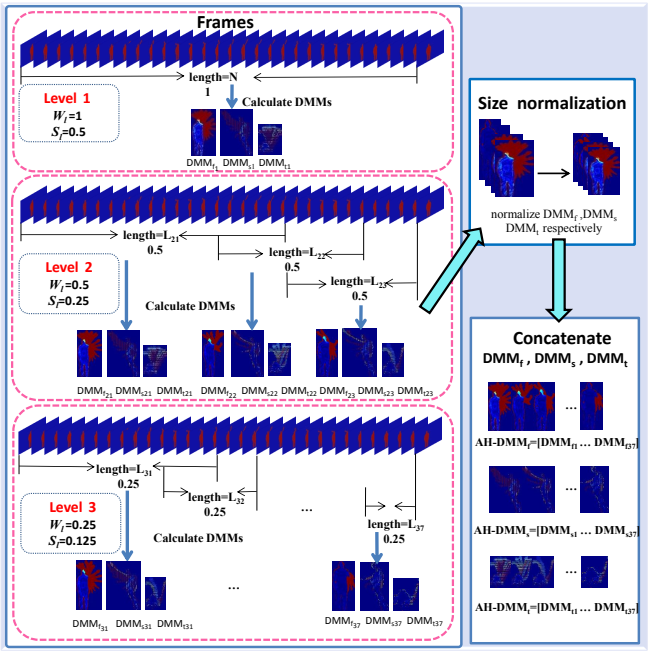


Fig. 2: The generation progress of AH-DMMs with three levels from a depth sequence. L_{lm} and DMM_{lm} refer to the frame length of m-th window and the m-th group DMMs in l-th level, respectively.

obversely, $ME(N)$=1. In level 1, DMMs are computed from entire sequence, and $W_1 = 1, S_1 = 0.5$. In level 2, DMMs are computed over three overlapping windows, corresponding to $W_2 = 0.5, S_2 = 0.25$. In level 3, the action sequence is divided into seven overlapping segments according to $W_3 = 0.25, S_3 = 0.125$, and we compute DMMs from each segment. Then DMM_f, DMM_s, DMM_t from all levels are normalized respectively and concatenated to form our AH-DMM_f, AH-DMM_s, AH-DMM_t.

Different from the temporal pyramid which divides sequences equally in time axis without overlapping [17], see Fig.1(b), the adaptive hierarchical structure divides sequences based on the distribution of motion energy. Therefore, it is insensitive to speed variations.

Compared with DMMs, the AH-DMMs encode temporal information of action sequences, more details of motion and more discriminative shape clues can be involved. Taking Fig.3 as an example, it illustrates the DMM_f and our AH-DMM_f of action "*draw tick*". It can be seen that the AH-DMM_f not only captures the information of a whole action, but also reflects the motion of sub-actions. From AH-DMM_f, we can observe the movement details of "*draw tick*" clearly. In addition, by using multi-size adaptive windows and steps, the intra-class variations caused by different action speeds can be reduced to a certain degree.

2.2. Gabor Filter Based Feature Representation

Gabor feature has shown effectiveness in capturing local structure and texture information of images and is popular in image processing field [18]. In this paper, Gabor filter

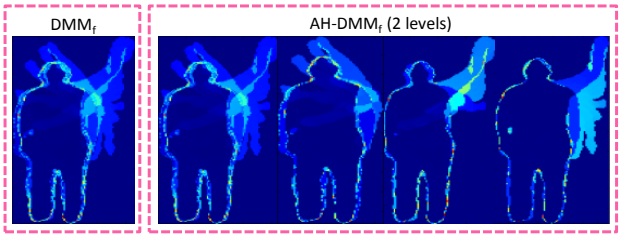


Fig. 3: Comparison between traditional $\mathrm{DMM_f}$ and $\mathrm{AH\text{-}DMM_f}$ of action "*draw tick*". $\mathrm{AH\text{-}DMM_f}$ captures the information of both the whole action and sub-actions and encodes more details of motion.

is chosen to characterize the local appearance and shape on AH-DMMs. We generate 40 Gabor filters with five scales and eight orientations, and then the Gabor features are computed by convolution of the AH-DMMs with Gabor filter. For front, side and top view, a d-dimensional Gabor feature vector can be obtained respectively. Then, we normalize the range of three feature vectors to [-1,1] for better classification accuracy and faster convergence rate. The three normalized feature vectors are denoted as $\mathbf{g}_{AH-DMM_f}$, $\mathbf{g}_{AH-DMM_s}$, $\mathbf{g}_{AH-DMM_t}$. The final feature representation $\mathbf{g}$ is the concatenation of three feature vectors, defined as follows:

$$\mathbf{g} = [\mathbf{g}_{AH-DMM_f}, \mathbf{g}_{AH-DMM_s}, \mathbf{g}_{AH-DMM_t}] \quad (4)$$

Since the concatenated feature vector has a high dimension, Principle Component Analysis (PCA) is adopted to reduce dimension and save 95% energy before classification.

2.3. Collaborative Representation Based Classification

The collaborative representation classifier (CRC) with l_2 -norm regularization has shown good classification performance and computational efficiency in face recognition [19], image classification [20] and action recognition [15], which motivates us to employ it in this paper.

Supposing that there are M training samples came from C classes of actions, each action sequence generates a feature vector $\mathbf{g}$ with d dimensions. The training set can be denoted as $\mathbf{G} = [\mathbf{G}_1, \mathbf{G}_2, ..., \mathbf{G}_C] \in R^{d \times M}$, where $\mathbf{G}_j = [\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_{m_j}]$ denotes m_j training samples from the j-th class, $j = 1, 2, ..., C$. Let $\mathbf{y} \in R^d$ denote a testing sample, the collaborative representation with l_2 -norm regularization can be mathematically represented as:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ \|\mathbf{y} - \mathbf{G}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_2^2 \right\} \quad (5)$$

where λ is the regularization parameter, $\boldsymbol{\alpha}$ is a coefficient vector corresponding to all the training samples. According to [19], the solution of CRC can be derived as:

$$\hat{\boldsymbol{\alpha}} = \left(\mathbf{G}^T\mathbf{G} + \lambda \cdot \mathbf{I}\right)^{-1} \mathbf{G}^T\mathbf{y} \quad (6)$$

After obtaining the coefficient vector, the residual errors between the feature vector $\mathbf{y}$ and the approximations can be calculated by:

$$r_j(\mathbf{y}) = \|\mathbf{y} - \mathbf{G}_j\hat{\boldsymbol{\alpha}}_j\|_2 \quad (7)$$

Table I: The performance of our AH-DMMs using different levels and comparison with the baseline TP-DMMs on MSRAction3D dataset under cross subject setting.

Levels	TP-DMMs	AH-DMMs	
	Accuracy(%)	Accuracy(%)	Time/Sequence(s)
l=1	90.11	90.11	0.42
l=2	90.91	**94.18**	1.56
l=3	92.36	93.45	3.81

here $\hat{\boldsymbol{\alpha}}_j$ is the coefficient vector associated with class j. Then the class label of $\mathbf{y}$ can be obtained as follows:

$$class(\mathbf{y}) = \arg\min_{j=1,...,C} r_j(\mathbf{y}) \quad (8)$$

3. EXPERIMENTS AND ANALYSIS

In this section, we evaluate the proposed method on benchmark MSRAction3D dataset [8] and DHA dataset [21]. We firstly describe the datasets and their respective evaluation setup, and then report our experimental results and analysis.

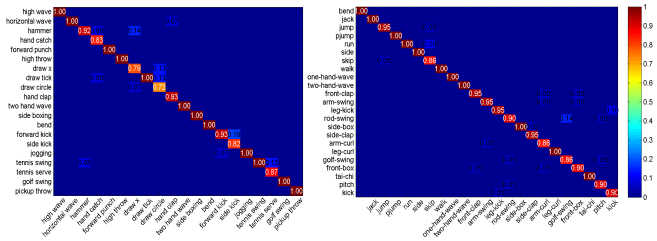
3.1. MSRAction3D Dataset

The MSRAction3D dataset contains 567 videos of 20 actions and each action is performed by 10 subjects for 2 or 3 times. This dataset is challenging for quite similar actions such as "*draw x*" and "*draw tick*", both of which have similar movements of hands.

Experimental settings. To ensure a fair comparison, we follow the cross subjects setting in [9], one half of the subjects (1, 3, 5, 7, 9) for training and the rest subjects (2, 4, 6, 8, 10) for testing. The size of each DMM_f, DMM_s, DMM_t are normalized to 102*52, 102*75, and 74*54 respectively, following [14]. The 5 scales and 8 orientations of gabor filter are $\nu \in \{0, 1, 2, 3, 4\}$, $\mu \in \{0, 1, 2, 3, 4, 5, 6, 7\}$, the size of filter template is fixed to $10 * 11$, whose values are chosen from [18]. The regularization parameter λ of CRC is assigned a value ranging from 0.0001 to 1, and we finally choose $\lambda = 0.001$, which leads to the highest recognition accuracy.

Evaluation of the hierarchical structure. The level l of our AH-DMMs is considered to have notable impact on the performance. Table I shows the recognition accuracies and computation time with different values of l from 1 to 3. It can be observed that our method obtains the best performance when l=2, and the highest accuracy of 94.18% is obtained.

Comparison with the baseline. To verify the validity of our method, it is compared with the baseline method: temporal pyramid based DMMs (TP-DMMs). The recognition accuracies are reported in Table I. As shown in Table I, when l=1, AH-DMMs and TP-DMMs have the same performance because both of them are equal to DMMs. While our method performs better than the baseline when l=2 and 3. It's because that TP-DMMs divides action sequence equally without overlapping, while our method utilizes motion energy-based segmentation strategy and is robust to intra-class variations.



(a) average accuracy =94.18% (b) average accuracy =95.45%

Fig. 4: Confusion matrices of our method for (a) MSRAction3D dataset and (b) DHA dataset

Table II: Comparison of the performance using different texture descriptors on AH-DMMs (l=2).

Texture descriptors	HOG	LBP	Gabor
Accuracy(%)	93.09	92.00	**94.18**

And AH-DMMs encode the information between two subsequences, so they capture more discriminative motion clues.

Evaluation of texture descriptors. We also compare the performance of AH-DMMs (l=2) when using different texture descriptors: Gabor, HOG and LBP. The experimental results are shown in Table II. It can be seen that Gabor descriptor performs better than HOG and LBP descriptors.

Comparison with the state-of-the-arts. Further we compare the performance of our method with several state-of-the-art methods on the MSRAction3D dataset and report the results in Table III. Particularly, we note that the accuracy of DMMs [13] is 88.73%, which is 5.45% lower than our method. The DMM-LBP [15] method which also utilizes depth motion maps obtains 93.00% accuracy, inferior 1.18% to ours. Besides, our method performs better than other methods using solo depth data such as "Random Occupancy Pattern" [9], "HOG3D+LLC" [22], "Super Normal Vector" [6], "Hierarchical 3D Kernel Descriptors" [23]. And it also outperforms the skeleton+depth based method "Moving Pose" [12]. Our method outperforms these methods mainly due to the following reasons. First, our AH-DMMs can sufficiently capture motion information of actions as well as shape clues of human body, and Gabor filter can effectively encode local structure and texture information of AH-DMMs; Second, with temporal information preserved, more details and more plentiful information of motion can be obtained; Third, our method is adaptive to speed variations, which helps to reduce intra-class variations. The confusion matrix is shown in Fig.4(a). It can be seen that 12 actions are 100% correctly recognized. The similar actions "*draw x*" and "*draw tick*" , "*horizontal wave*" and "*high wave*" are distinguished successfully.

Table III: Comparison with the state-of-the-arts on MSRAction3D dataset under cross subject setting.

Approaches	Year	Accuracy(%)
Bag of 3D Points [8]	2010	74.70
DMM-HOG [13]	2012	88.73
Random Occupancy Pattern [9]	2012	86.50
HON4D [7]	2013	88.89
DSTIP [10]	2013	89.30
HOG3D+LLC [22]	2016	90.90
Moving Pose [12]	2012	91.70
DMM-LBP [15]	2015	93.00
Super Normal Vector [6]	2014	93.09
Hierarchical 3D Kernel Descriptors [23]	2016	93.99
AH-DMMs+Gabor (Ours)	2016	**94.18**

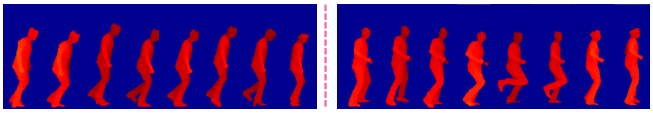


Fig. 5: Similar lateral actions, such as "*skip*" (left) and "*run*" (right) in DHA dataset.

Table IV: Comparison with the state-of-the-arts on DHA dataset under LOSO setting.

Approaches	Year	Accuarcy (%)
DMHI-Gist [24]	2014	86.00
D-STV [21]	2012	86.80
SDM-BSM [25]	2015	89.95
DMPP-PHOG [26]	2015	95.00
AH-DMMs+Gabor (Ours)	2016	**95.45**

3.2. DHA Dataset

DHA dataset contains 483 videos of 23 different actions, and each action is performed by 21 actors once. This dataset is challenging because its inter-class ambiguity is quite large. The parameter settings are the same as MSRAction 3D dataset and the evaluation setting adopted in DHA dataset is Leave-One-Subject-Out (LOSO) setting [25].

The confusion matrix is shown in Fig.4(b). It shows that 11 actions are 100% correctly recognized and 17 actions reach 95% recognition accuracy. The classification error occurs in quite easily confused actions such as "*skip*" and "*run*" due to their similar lateral motion patterns (as shown in Fig. 5). Table IV shows the experimental results compared with the state-of-the-art methods on DHA dataset. It can be seen that our method achieves the highest recognition accuracy of 95.45% in this dataset, which also proves the effectiveness and robustness of the proposed method.

4. CONCLUSIONS

This paper presents an effective method for human action recognition using AH-DMMs and Gabor filter. The proposed AH-DMMs can capture more details of motion and shape clues by preserving the temporal information of actions. Meanwhile, the AH-DMMs are adaptive to action speed variations for using energy-based hierarchical structure. Gabor filter is then adopted to encode texture information of AH-DMMs and generates more compact action representations, and CRC are utilized as the classifier. The experimental results on two benchmark datasets show that our method outperforms the state-of-the-art approaches. In future work, we will focus on combining skeleton joints with depth data to further improve the recognition accuracy.

5. REFERENCES

[1] M. Liu and H. Liu, "Depth context: a new descriptor for human activity recognition by using sole depth sequences," *Elsevier Neurocomputing*, vol. 175, pp. 747–758, 2016.

[2] I. Rodomagoulakis, N. Kardaris, and V. Pitsikalis, "Multimodal human action recognition in assistive human-robot interaction," *in ICASSP*, pp. 2702–2706, 2016.

[3] W. Lin, M. Sun, R. Poovendran, and Z. Zhang, "Activity recognition using a combination of category components and local models for video surveillance," *IEEE TCSVT*, vol. 18, no. 8, pp. 1128–1139, 2008.

[4] M. Liu, H. Liu, and Q. Sun, "Action classification by exploring directional co-occurrence of weighted stips," *in ICIP*, pp. 1460–1464, 2014.

[5] H. Liu, M. Liu, and Q. Sun, "Learning directional co-occurrence for human action classification," *in ICASSP*, pp. 1235–1239, 2014.

[6] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," *in CVPR*, pp. 804–811, 2014.

[7] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," *in CVPR*, pp. 716–723, 2013.

[8] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," *in CVPRW*, pp. 9–14, 2010.

[9] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," *in ECCV*, pp. 872–885, 2012.

[10] L. Xia and J. K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," *in CVPR*, pp. 2834–2841, 2013.

[11] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," *in CVPRW*, pp. 14–19, 2012.

[12] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," *in ICCV*, pp. 2752–2759, 2013.

[13] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," *in ACM Multimedia*, pp. 1057–1060, 2012.

[14] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *Springer JRTIP*, pp. 1–9, 2013.

[15] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," *in WACV*, pp. 1092–1099, 2015.

[16] P. Wang, W. Li, Z. Gao, C. Tang, and P. Zhang, J.and Ogunbona, "Convnets-based action recognition from depth maps through virtual cameras and pseudo-coloring," *in ACM Multimedia*, pp. 1119–1122, 2015.

[17] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," *in CVPR*, pp. 2847–2854, 2012.

[18] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE TIP*, vol. 11, no. 4, pp. 467–476, 2002.

[19] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?," *in ICCV*, pp. 471–478, 2011.

[20] M. Yang, L. Zhang, D. Zhang, and S.L. Wang, "Relaxed collaborative representation for pattern classification," *in CVPR*, pp. 2224–2231, 2012.

[21] Y. Lin, M. Hu, W. Cheng, Y. Hsieh, and H. Chen, "Human action recognition and retrieval using sole depth information," *in ACM Multimedia*, pp. 1053–1056, 2012.

[22] H. Rahmani, D.Q. Huynh, A. Mahmood, and A. Mian, "Discriminative human action classification using locality-constrained linear coding," *Elsevier PRL*, vol. 72, pp. 62–71, 2016.

[23] Y. Kong, B. Satarboroujeni, and Y. Fu, "Learning hierarchical 3d kernel descriptors for rgb-d action recognition," *Elsevier CVIU*, vol. 144, pp. 14–23, 2016.

[24] Z. Gao, J. Song, H. Zhang, A. Liu, Y. Xue, and G. Xu, "Human action recognition via multi-modality information," *JEET*, vol. 9, no. 2, pp. 739–748, 2014.

[25] H. Liu, L. Tian, M. Liu, and H. Tang, "Sdm-bsm: A fusing depth scheme for human action recognition," *in ICIP*, pp. 4674–4678, 2015.

[26] Z. Gao, H Zhang, G.P. Xu, and Y.B. Xue, "Multi-perspective and multi-modality joint representation and recognition model for 3d action recognition," *Elsevier Neurocomputing*, vol. 151, pp. 554–564, 2015.